Chemistry Education Research and Practice

ROYAL SOCIETY OF CHEMISTRY

COMMENT

View Article Online
View Journal | View Issue

Cite this: *Chem. Educ. Res. Pract.*, 2020, **21**, 1218

Comment on "Increasing chemistry students' knowledge, confidence, and conceptual understanding of pH using a collaborative computer pH simulation" by S. W. Watson, A. V. Dubrovskiy and M. L. Peters, *Chem. Educ. Res. Pract.*, 2020, 21, 528

Keith S. Taber

Received 28th April 2020,
Accepted 12th June 2020

DOI: 10.1039/d0rp00131g

rsc.li/cerp

This comment discusses some issues about the use and reporting of experimental studies in education, illustrated by a recently published study that claimed (i) that an educational innovation was effective despite outcomes not reaching statistical significance, and (ii) that this refuted the findings of an earlier study. The two key issues raised concern how the research community should understand the concept of refutation when comparing across studies, and whether the adoption of inferential statistics in a study should bind researchers to accept the inferences such tests suggest.

Introduction

A recent paper in CERP (Watson *et al.*, 2020) reported an investigation of the value of collaborative use of a computer simulation (a 'PhET sim') in undergraduate learning. This was a quasi-experimental study, and because it was not possible to meet the conditions for a true experiment (such as randomisation of participants to conditions) the researchers made pre-test comparisons between those learners in the innovation and comparison conditions.† This is a well-motivated and interesting study, and it is reported in sufficient detail for a reader to evaluate, and acknowledges that the "study is best interpreted with a consideration for its limitations" (p. 534).

This comment raises some issues about the interpretation and presentation of results in this study – issues that actually arise quite widely in experimental research in education (Taber, 2019). The authors conclude from their study that "the findings of this research study indicated that the collaborative PhET sim positively impacted students' pH conceptual understandings, which both supports and refutes the literature" (p. 534), but it can be questioned whether this should be concluded given the reported results. More specifically, comments are offered regarding two statements found in the Discussion section of the paper (p. 533):

(i) "This research study found that collaborative computer sim group members experienced higher mean scores regarding pH knowledge and conceptual understanding, and indicated higher levels of pH-related confidence from the beginning to the end of the semester when compared to the traditional group members";

(ii) "our findings refute those of Hawkins and Phelps (2013) who found no statistical difference in learning gains between a group of students using a computerized sim on electrochemistry (treatment) *versus* a group of students who were taught electrochemistry *via* a traditional hands-on experiment (control)".

Implications of adopting statistical tests

It is common practice when reporting the results of experimental studies, such as those into the effectiveness of a teaching innovation, to not only report descriptive statistics (*e.g.*, mean outcomes on the measures made), but to also use statistical tests to compare between conditions to see if differences in outcomes reach statistical significance. In their recent study, Watson and colleagues report both mean values for various measurements taken, and the outcomes of statistical tests. They report a number of comparisons between the pre-learning and post-learning

Faculty of Education, University of Cambridge, UK. E-mail: kst24@cam.ac.uk

† Watson and colleagues refer to the experimental "computer sim group/intervention" condition, and the "traditional/control group". As well as simplifying language, a reason for preferring different terms here is that arguably both conditions were interventions (*i.e.*, activities intended to bring about learning), and – as the authors acknowledge (p. 534) – there was no genuine control condition: for example, the two groups may have spent different amount of time working with the materials. The term comparison condition is widely used when there is no strict control condition.

Table 1 Comparisons between pre-test and post-test

What is compared from pre-test to post-test	Outcome
Knowledge of pH (all participants)	n.s. ($p = 0.419$)
Confidence in understanding of pH (all participants)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Confidence in understanding of pH (comparison group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (all participants)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (innovation group)	Significant increase ($p < 0.001$)
Conceptual understanding of pH (comparison group)	Significant increase [no p value cited]

Table 2 Comparisons between the innovation and comparison conditions

What is compared between conditions	Outcome
Knowledge before studying	n.s. ($p = 0.712$) [seen as evidence of equivalence]
Knowledge after studying	n.s. ($p = 0.460$)
Increase in confidence in understanding	n.s. [no p value cited]
Understanding before studying	n.s. ($p = 0.384$) [seen as evidence of equivalence]
Understanding after studying	n.s. ($p = 0.068$)

measurements (see my Table 1), and between the two conditions (see my Table 2).

In abstracting results for my tables I have for present purposes only noted whether or not differences found were considered statistically significant (*i.e.*, $p < 0.05$; n.s. = not significant), and, where these are reported in the paper, the p values that such judgements were based on. Table 1 shows that whilst test scores intended to measure knowledge of pH do not show sufficient increases to be found statistically significant, there were statistically significant increases across two other measures: confidence in understanding of pH, and conceptual understanding of pH.

The results in Table 1 show that after the learning activities there were significant increases in these two measures across both conditions. Tests of significance by themselves offer limited information, and it is often recommended that effect sizes should also be calculated to give an indication of the likely practical importance of any differences found (Sullivan and Feinn, 2012). Watson and colleagues report having calculated effect sizes, but only inform readers of their results for two of these calculations.

These increases in measurements after learning suggest that the teaching input likely had an effect on both conceptual understanding and student confidence as these increases were sufficiently unlikely to occur by chance to be judged 'significant'. Such conclusions are probabilistic and subject to assumptions (*e.g.*, that the measuring instruments provided valid measurements, and that there were no unidentified confounding variables – such as students independently undertaking study of the same topic in another course). It is reasonable to infer that working with the PhET simulation probably led to improvements in student understanding and confidence – however the results in Table 1 also show the same is probably true of the comparison input, and so the results reproduced in Table 1 suggest both instructional inputs seem to have an effect, but do not allow any conclusion about whether one or other input was more effective.

The results abstracted into Table 2 are those which seek to compare between the two student groups in the different conditions. As Watson and colleagues are employing inferential statistics as the basis of their experimental study, Table 2 presents outcomes in terms of statistical significance, and shows each result was non-significant. Those entries in Table 2 relating to pre-test measurements are concerned with testing for equivalence between groups. When it is not possible to randomise participants to conditions it is quite feasible that systematic differences between the groups at the outset may be sufficient in themselves to lead to substantially different post-test results.‡ Equivalence tests are intended to show groups are initially sufficiently similar for any differences between them at that point to be unlikely to be responsible for any significant post-test differences.

Simply demonstrating that initial differences do not reach statistical significance (*i.e.*, at a level that would only occur by chance with very low probability) is a very limited test of equivalence (Taber, 2013). Any choice of p value is somewhat arbitrary, but it has been suggested that if using p for this purpose then a substantially higher value of p, $p > 0.5$ (not reached here when comparing conceptual understanding at pre-test), should be adopted as a suitable criterion for equivalence (Taber, 2019) – simply because it at least answers the nominal question "are the measured differences between the groups likely to be due to chance variation rather than systematic differences?" with "probably". That would still be a simplistic treatment. Given how p values are sensitive to sample sizes, more sophisticated analyses have been recommended (see, for example, Lewis and Lewis, 2005), and these may involve the subject experts making a judgment of what would be a large enough difference in scores to make a practical difference, before actually carrying out the pre-testing. Yet, whilst Watson and colleagues are not using a

‡ This is still possible when randomisation occurs, but has a low probability, which is acknowledged in that a statistically significant difference at post-test is considered to have a low (not but zero) probability of being due to chance events.

robust test of equivalence in their study, it should be noted they are following common practice.

Of particular note, however, are those comparisons between the students in the two conditions at post-test. As shown in Table 2, these did not reach significance. No statistically significant difference in measured outcomes was found between the innovative learning condition using a simulation and the comparison condition based on reading assignments.§ Yet, as we have seen, the authors conclude that students in the innovation condition had ''higher mean scores regarding pH knowledge and conceptual understanding, and indicated higher levels of [increases in] pH-related confidence'' and that ''the collaborative PhET sim positively impacted students' pH conceptual understandings''.

There is no logical contradiction between higher mean scores and non-significant differences. However, Watson and colleagues appear to be setting aside an important scientific norm here. In experimental research it is important to set out in advance both a clear hypothesis to be tested (which they have done), and the criterion (or criteria) that will be used to reach conclusions. The use and reporting of inferential statistics indicates that tests of significance will be adopted as the basis for inferring whether an experiment has produced a positive outcome. Here, Watson and colleagues have indeed used such statistics as the grounds for claiming they have baseline equivalence at pre-test, but they then set aside the results of those tests to claim positive outcomes at post-test. This seems to be the application of a double standard.

Reporting context-directed research

Anyone who has carried out practitioner research may have sympathy here with Watson and colleagues. It may often seem to those involved in a study that differences in outcomes that are non-significant, *i.e.*, not sufficiently statistically unlikely to reach significance, may still be of practical importance (and indeed effect sizes may seem more informative here). It may be useful to consider a distinction that has been made between theory-directed and context-directed research (Taber, 2013). In theory-directed research, the main motivation for a study is to examine a generalisable theoretical question (such as about the efficacy of PhET simulations in teaching chemistry topics) and a choice of research site(s) and sample is then made as an instrumental means of obtaining findings that can be generalised to a wide range of contexts. In such studies it would seem perverse to draw conclusions contrary to the outcomes of the inferential tests undertaken.

By contrast, much research in education is context-directed: it is primarily intended to better understand, and improve, teaching and learning at a research site where the investigators have more than an instrumental interest – often their own teaching context and students. In evaluating the outcomes of such research it would seem contrary to dismiss measured differences that could have educational importance just because they do not reach statistical significance (especially if the p value shows a result is statistically unlikely, whilst not reaching the strict cut-off of 0.05 – as is the case for the measurement of student conceptual understanding in Table 2). For practitioners, it may be logical not to discard an innovation that was measured to have a positive effect but not to be significantly significant. At the very least, it may make sense to repeat the experiment with a subsequent cohort to build up the evidence base.

So, it is suggested that the implications of a study may be practically different when we ask (a) whether it is sensible for the investigators to persevere with an innovation in the same teaching and learning context in future practice, and (b) whether the experimental outcomes are strong enough to draw generalised conclusions that justify recommending changes of practice elsewhere. In the study by Watson and colleagues, there would seem to be strong encouragement for them to sensibly decide to proceed with the use of the PhET simulation in their teaching, and to continue to monitor its effectiveness. However, the usual conventions of educational research suggest their findings are not (yet, at least) strong enough to conclude recommending wide-scale implementation of the innovation more generally.

What do we mean by refutation?

The other statement which invites comment concerns refutation. Watson and colleges suggest that their findings ''are in agreement'' with previous studies that had ''concluded that when students engage with supplemental multi-media to enhance traditional lecture instruction, learning gains are realized''¶ but refute the findings of Hawkins and Phelps (p. 533). This raises a question of how refutation is to be understood.

The kind of entity that can be refuted is a statement or claim. Hawkins and Phelps had made a claim about the outcomes of an educational experiment carried out in a particular teaching and learning context. If an attempt had been made to replicate their study in the same context (say, with the subsequent cohort of students) then it is possible there would have been a different result. Yet, even so, that would not refute the claim made about the findings of the original study undertaken with the first sample of learners. Unlike in the physical sciences, where the findings of research on samples/specimens of natural kinds (*e.g.*, graphite rods, benzene) can be assumed to apply to other samples/specimens of the same kinds, generalisation is more problematic in education (Lederman, 2020; Taber, 2019, 2020).

§ For conceptual understanding the p value (0.068) was close to reaching significance – however, if we accept the principle that in an experiment we need to decide on a criterion for significance before undertaking an analysis, then strictly we need to treat any 'almost statistically significant' result as non-significant. Readers are told that in this study ''a significance value of 0.05 was utilized'' (p. 532).

¶ Watson *et al.*'s study design did not involve providing ''supplemental'' experience to ''to enhance traditional'' teaching, but (like the research of Hawkins and Phelps) substituted one learning experience for another.

If the same simulation had been used to learn about the same topic with a sample of learners in a different institutional context, then comparing outcomes between studies would speak to issues of generalisability across contexts, rather than confirmation/refutation of the earlier study. It has been argued that generalisability in studies of teaching innovations needs to be based on exploring the extent of the domains of learning contexts where particular pedagogic innovations can be found to be effective (Taber, 2019). That is, studies of teaching approaches, or learning resources, undertaken in different educational contexts should not be seen to confirm or refute each other's specific findings, but rather to address issues of the range of conditions where something might be expected to work.

There are other substantive differences between the two studies. Whether a study comparing learning from a simulation about pH with ''traditional reading-based classroom assignments'' (p. 531) could, in principle, refute the findings of a study comparing learning from an ''electrochemistry virtual lab simulation'' with a ''hands-on experiment'' (Hawkins and Phelps, 2013, p. 517) is questionable. The more recent study's findings could certainly potentially conflict with any generalised conclusion (*e.g.*, that simulations are always/never effective learning tools), had the authors of the earlier study been bold (or cavalier) enough to make such a claim, but Hawkins and Phelps actually concluded that ''more research needs to be done to determine virtual laboratories efficacy as a replacement for more traditional hands-on laboratory experiences'' (p. 521).

The double standard referred to above would in any case cast doubts over Watson and colleagues' claim to have refuted the earlier study, even if this was accepted as an 'in principle' possibility. There is surely a 'sleight of hand' in first eschewing the significance criterion in making their own claim for positive outcomes, and then using this conclusion as the basis for seeking to refute the results of other researchers ''who found no statistical difference in learning gains''. There would seem to be a logical flaw in a claim that a study that finds no significant difference in learning gains can refute another study that also found no significant difference in learning gains.

Recommendations

My intention here is not to criticise a particular study, which as I noted at the outset has considerable merit. Rather, I think Watson and colleagues' work raises issues about the way we apply ideas about experimental methodology deriving originally from laboratory work in the natural sciences to teaching and learning contexts, and highlights the need for the research community to give more critical attention to the challenges of educational experiments and how they are best conceptualised and reported. The discussion in this comment leads me to the following recommendations:

(1) When it is not possible to randomise study participants to different conditions, a stronger demonstration of equivalence should be adopted than just checking for a statistically significant pre-test difference between groups: in particular, making a judgment of what would be a large enough difference in scores to make a practical difference, before actually carrying out the pre-testing.

(2) Where possible, effect sizes should be calculated to offer a measure of how substantive differences in outcomes between groups are, and these should be provided alongside the results of tests of statistical significance.

(3) Once a confidence level has been adopted for use in statistical tests (*e.g.*, $p < 0.05$), all conclusions drawn should be consistent with this level: *i.e.*, positive outcomes should not be claimed where $p \geq 0.05$. Where authors feel that a non-significant difference is substantial enough to inform their future practice, they should be careful to distinguish this pragmatic stance, from the formal outcomes of the experiment.

(4) Authors should avoid suggesting they have refuted previous studies (unless a good case can be made that there is a close replication of the original study) and rather frame their comparisons in terms of the extent to which findings between studies support the same generalised conclusions.

(5) Authors should offer sufficient description of the context of their particular studies to allow reports to contribute to incrementally developing guidance on the ranges of application of different types of innovation.

Conflicts of interest

There are no conflicts to declare.

References

Hawkins I. and Phelps A. J., (2013), Virtual laboratory vs. traditional laboratory: which is more effective for teaching electrochemistry? *Chem. Educ. Res. Pract.*, **14**(4), 516–523, DOI: 10.1039/C3RP00070B.

Lederman N., (2020), Replicable, reproducible, and generalisable: Implications of scientific hallmarks for research in education. *HPS&ST Newsletter*, April, 6–13.

Lewis S. E. and Lewis J. E., (2005), The same or not the same: Equivalence as an issue in educational research, *J. Chem. Educ.*, **82**(9), 1408, DOI: 10.1021/ed082p1408.

Sullivan G. M. and Feinn R., (2012), Using effect size—or why the *p* value is not enough, *J. Grad. Med. Educ.*, **4**(3), 279–282, DOI: 10.4300/JGME-D-12-00156.1.

Taber K. S., (2013), *Classroom-based Research and Evidence-based Practice: An introduction*, 2nd edn, London, Sage.

Taber K. S., (2019), Experimental research into teaching innovations: responding to methodological and ethical challenges, *Stud. Sci. Educ.*, **55**(1), 69–119, DOI: 10.1080/03057267.2019.1658058.

Taber K. S., (2020), Is reproducibility a realistic norm for scientific research into teaching? *HPS&ST Newsletter*, April, 13–23.

Watson S. W., Dubrovskiy A. V. and Peters M. L., (2020), Increasing chemistry students' knowledge, confidence, and conceptual understanding of pH using a collaborative computer pH simulation, *Chem. Educ. Res. Pract.*, **21**(2), 528–535, DOI: 10.1039/C9RP00235A.